Convergence Rate Analysis of Distributed Gossip (Linear Parameter) Estimation: Fundamental Limits and Tradeoffs

Soummya Kar and José M. F. Moura*

Abstract

The paper considers gossip distributed estimation of a (static) distributed random field (a.k.a., large scale unknown parameter vector) observed by sparsely interconnected sensors, each of which only observes a small fraction of the field. We consider linear distributed estimators whose structure combines the information *flow* among sensors (the *consensus* term resulting from the local gossiping exchange among sensors when they are able to communicate) and the information *gathering* measured by the sensors (the *sensing* or *innovations* term.) This leads to mixed time scale algorithms–one time scale associated with the consensus and the other with the innovations. The paper establishes a distributed observability condition (global observability plus mean connectedness) under which the distributed estimates are consistent and asymptotically normal. We introduce the distributed notion equivalent to the (centralized) Fisher information rate, which is a bound on the mean square error reduction rate of any distributed estimator; we show that under the appropriate modeling and structural network communication conditions (gossip protocol) the distributed gossip estimator attains this distributed Fisher information rate, asymptotically achieving the performance of the optimal centralized estimator. Finally, we study the behavior of the distributed gossip estimator when the measurements fade (noise variance grows) with time; in particular, we consider the maximum rate at which the noise variance can grow and still the distributed estimator being consistent, by showing that, as long as the centralized estimator is consistent, the distributed estimator remains consistent.

Keywords: Distributed estimation, gossip, random networks, sensor networks, link failures, switching topology

The first author is with the Dep. Electrical Engineering, Princeton University, Princeton, NJ. This work was performed while the first author was with the Dep. Electrical and Computer Engineering, Carnegie Mellon University. The second author is with the Dep. Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA (e-mail: soummyak@andrew.cmu.edu, moura@ece.cmu.edu, ph: (412)268-6341, fax: (412)268-3890.)

Work partially supported by AFOSR grant # FA95501010291; and by NSF grant # CCF1011903.

I. INTRODUCTION

A. Motivation

We consider distributed (or decentralized) estimation of a random field where observations are collected by possibly a large number of sparsely internetworked sensors. The network operates under the gossip random protocol and may be subject to random infrastructure failures (communication channels may fail intermittently.) There is no fusion-center and the estimation is performed locally at each sensor with inter-sensor message exchanges occurring at random times. Because the random field of interest is distributed, each sensor can only observe a part of the field, and no sensor can in isolation obtain a reasonable estimate of the entire field. This paper studies the conditions under which the distributed algorithms operating under the random intermittent conditions (gossip and link failures) that we consider can achieve (asymptotically) performance that is equivalent to the estimation performance of centralized optimal algorithms. To be more concrete and as an abstraction of the environment[1], we model it by a static vector parameter, whose dimension, M, can be arbitrarily large. Each sensor's observations, say for sensor n, are M_n dimensional noisy measurements of a *part* of the (static random) field, where $M_n \ll M$. We assume that the sensing rate, i.e., rate of receiving observations at each sensor, is comparable to the communication rate among sensors, so that sensors update their estimate at time index i by fusing appropriately their current estimate with the observation (innovation) at i and the estimates at i received from those sensors with which it successfully gossips at i. Because of the communication intermittency, the distributed estimators that we consider exhibit mixed time scales: one associated with the *consensus*, i.e., mixing estimation updating resulting from receiving the estimates from the neighbors; and the other associated with the *sensing* or estimation updating from the *innovations*. In this paper, we consider a general class of *linear* distributed gossip networked estimators and study the conditions under which they exhibit the same estimation error convergence rate as a centralized linear field estimator. Nonlinear distributed estimators and distributed estimation of time varying random fields under the gossip protocol are considered elsewhere, [1] and [2], respectively.

We discuss the major challenges in gossip distributed estimation and highlight the key contributions of the paper:

- **Infrastructure failures and gossip communication**: The inter-sensor communication may be bandwidth and power constrained and subject to random environmental conditions. For example, the sensors may share a common wireless medium and, due to competing objectives, the inter-sensor transmissions may be scheduled by the underlying MAC (Medium Access Control) layer to occur at random times; in fact, in many situations of interest, the exact medium access (MAC) protocol

[1]The term environment or field has a generic usage here. It may correspond to sensors deployed over a domain of interest like a temperature surface, or, a networked physical system instrumented with sensors. Typical examples of the latter include cyberphysical systems like the power grid, and networked control systems (NCS), where a network of distributed actuators are equipped with sensors.

(randomized) is not known or determined àpriori, the inter-sensor communications is asynchronous, and random data packet dropouts may occur.

- **Distributed observability**: It is well known that centralized estimation requires observability conditions to be satisfied for the estimation task to be *successful*[2]. As we will see, formulating a satisfactory notion of *distributed observability* is not trivial. A difficulty stems from the distributed nature of the *information*, i.e., sensors observe only a portion of the field of interest. The incorporation of estimate fusion among the sensor nodes (*consensus*) together with local innovation updates suggest that distributed observability should be not only a function of the sensor observations, but closely tied to the structural properties of the communication network governing the information flow. These conditions are sensitive to the pattern of information dissemination in the network and depends on the level of node cooperation, for example, gossiping. We present minimal conditions for distributed observability, namely, for example, in the case of full cooperation (each node exchanges its entire estimate with its neighbors), we show that *global observability*[3] and mean connectedness of the time varying communication graph are sufficient to ensure *consistent* parameter estimates at *each* sensor.
- **Distributed versus optimal centralized estimation**: We show that under reasonable assumptions, the gossip distributed estimators we develop, like the centralized optimal estimator, lead to consistent parameter estimates at each sensor. The natural question of interest is to compare the rate of convergence of these schemes to the true parameter value. We adopt asymptotic normality and the associated asymptotic variance as the metric for comparing different estimators. It is known from the theory of recursive estimation (centralized), that the optimum centralized estimator (under reasonable assumptions) achieves asymptotic variance equal to the Fisher information rate. In this paper, we formalize a notion of distributed Fisher information rate, i.e., a lower bound on the asymptotic variance of all distributed schemes and also investigate the existence of optimal distributed estimators achieving this lower bound. It turns out that, if the inter-sensor communication is noisy or quantized, the asymptotic variance of distributed estimators is always higher than their centralized counterpart. On the other hand, a remarkable asymptotic time scale separation phenomenon shows that, in the absence of channel noise or quantization (but presence of random link failures and gossip,) there exist distributed estimation schemes whose asymptotic variance equals the centralized Fisher information rate under pragmatic conditions. In particular, it is shown that, in a Gaussian environment, a distributed estimator is equivalent to a centralized one in terms of asymptotic variance, and, more generally, equivalent to the best linear centralized estimator. This is significant, as it shows that, under reasonable assumptions, a distributed gossip estimator is as good as a centralized one, the latter having access to all sensor observations at all times. We present some intuitive remarks. In a

[2]Successful means the estimate sequence generated over time possesses desirable properties like consistency, asymptotic normality etc.

[3]Global observability corresponds to the centralized setting, where an estimator has access to the observations of all sensors at all times. The assumption of global observability does not mean that each sensor is observable; rather, that if there was a centralized estimator with simultaneous access to all the sensor measurements, this centralized estimator would be observable.

centralized recursive (parameter) estimation scheme, the estimate update rule involves combining the past estimate with the new innovation (observation), the key design parameter being the time varying gain or weight associated to the innovation term. Since, the observations are noisy, for parameter estimation, this weight sequence needs to go to zero for achieving convergence and, in fact, needs to be square summable to constrain the effect of the observation noise. In most cases, assuming independent observations over time, the innovation gains decrease as $1/i$ (i being the iteration or time index) for optimal estimation performance. This means that the estimation uncertainty cannot be reduced at a rate $1/\sqrt{i}$, a consequence of central limit theorem type arguments. Now, consider the distributed scheme. Here, the algorithm design involves two gain sequences, one for the local innovations at each sensor and the other for estimate fusion (consensus) across sensors. To design good performance distributed gossip estimators, the trick is in choosing the fusion or consensus gain properly, so that its effect decays at a slower rate than the innovation gain. In the absence of quantization or channel noise, it is possible to choose the consensus weight sequence such that its squared sum goes to ∞, in contrast to the innovation weight sequence whose squared sum needs to be finite. It is shown that this tuning of the different gain sequences leads to an asymptotic time scale separation, the rate of information dissemination dominating the rate of reduction of uncertainty by observation acquisition. This tuning is not possible in the case of quantized or noisy transmissions, as each consensus step introduces noise, preventing proper adjustment of the gain sequences. The analysis approach that we develop is of independent interest and contributes to the theory of mixed time scale stochastic approximation.[4] Related to our mixed time scale algorithms is the work [4], which develops methods to analyze such algorithms in the context of simulated annealing. In [4] the role of our innovation potential is played by a martingale difference term. However, in our paper, an additional difficulty with respect to [4] is that the innovation is not a martingale difference process, and so a key step in our analysis is to derive pathwise strong approximation results to characterize the rate at which the innovation process converges to a martingale difference process.

Brief review of the literature. We comment on the relevant literature. An early treatment of distributed stochastic algorithms appears in [5] (see also [6], [7], [8].) In [5], almost sure convergence is established for a class of distributed stochastic algorithms in the context of distributed optimization. This line of work assumes the existence of a fixed time window T, such that the union of communication graphs over any interval of length T is connected with probability one. Also, the stochastic noise appears only in the computation of the local gradients that play the role of innovations in our approach. The conditions imposed on the local gradients are rather strong and implicitly assume that the individual processor (sensor in our terminology) dynamics are stable. Some of these conditions are relaxed in [8], which derives almost sure convergence and asymptotic normality for a class of constrained and unconstrained

[4]By mixed time scale, we refer to stochastic algorithms where two potentials act in the same update step with different weight or gain sequences. This should not be confused with stochastic algorithms with coupling (see [3]), where a quickly switching parameter influences the relatively slower dynamics of another state, leading to *averaged* dynamics.

parallel and communicating stochastic procedures with perfect communication. On the contrary, the gossip distributed estimators we develop in this paper are general mixed time-scale procedures in generic random environments and provide pathwise strong convergence rates. Our work does not impose local conditions on the innovation processes and develops and infers connective stability based on structural network conditions and global observability and establishes strong invariance results relating network information flow and the effect of local innovations.

More recently, there has been renewed interest in distributed approaches motivated by wireless sensor networks (WSN) applications. The papers [9], [10], [11], [12] study the estimation problem in static networks, where either the sensors take a single snapshot of the field at the start and then initiate distributed consensus protocols (or more generally distributed optimization, as in [10]) to fuse the initial estimates, or the observation rate of the sensors is assumed to be much slower than the inter-sensor communicate rate, thus permitting a separation of the two time-scales. More relevant to our work are [13], [14], [15], [16], which consider the linear estimation problem in non-random networks, where the observation and consensus protocols are incorporated in the same iteration. In [13], [15], the distributed linear estimation problems are treated in the context of distributed least-mean-square (LMS) filtering, where constant weight sequences are used to prove mean-square stability of the filter. The use of non-decaying combining weights in [13], [15], [16] leads to a residual error; however, under appropriate assumptions, these algorithms can be adapted for tracking certain time-varying parameters. The distributed LMS algorithm in [14] considers decaying weight sequences, thereby establishing $\mathcal{L}_2$ convergence to the true parameter value. In contrast to these, our work quantifies the pathwise information dissemination rate and its relation to the innovation rate by studying general mixed time-scale procedures. We consider structural conditions based on the network topology and observation pattern to develop a satisfactory notion of distributed observability and provide fundamental limits on the performance of distributed schemes.

The key difference between the current paper and the linear algorithm $\mathcal{LU}$ in [1] involves the use of different weight sequences for the consensus and the innovation terms, giving to the linear distributed estimators here a mixed time scale behavior. On the other hand, in this paper, we assume unquantized transmissions in the distributed gossip estimators. Another difference that will be noted below is the incorporation of a general matrix gain K into the innovation update. These modifications make the technical analysis of the distributed gossip linear estimators in this paper highly non-trivial and very distinct from the analysis of $\mathcal{LU}$ in [1].

We briefly comment on the organization of the rest of the paper. Section I-B sets up notation and preliminary concepts to be used throughout the paper. Section II formulates the distributed estimation problem, introduces the algorithm $\mathcal{GLU}$ and the assumptions (Section II-A.) Some technical results on the convergence of stochastic recurrences are established in Section III. This section also considers some properties of centralized estimators, with which we compare our distributed scheme. The main results of the paper are stated in Section IV. Section V develops convergence properties of the $\mathcal{GLU}$ algorithm, leading to the proofs of the main theorems in Section VI. Finally, Section VII concludes the paper.

B. Notation

We denote the k-dimensional Euclidean space by $\mathbb{R}^k$. The set of $m \times n$ matrices with real entries is denoted by $\mathbb{R}^{m\times n}$. $\mathbb{S}^N, \mathbb{S}^N_+, \mathbb{S}^N_{++}$ refer to the subsets of symmetric, positive semidefinite, positive definite matrices in $\mathbb{R}^{N\times N}$ respectively. The $k \times k$ identity matrix is denoted by I_k, while $\mathbf{1}_k, \mathbf{0}_k$ denote respectively the column vector of ones and zeros in $\mathbb{R}^k$. The set of integers is denoted by $\mathbb{T}$, whereas $\mathbb{N}$ stands for the natural numbers. $\mathbb{T}_+$ denotes the set of nonnegative integers and indices the iteration time slots throughout the paper.

Define the rank one $k \times k$ matrix P_k by

$$P_k = \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^T \tag{1}$$

The only non-zero eigenvalue of P_k is one, and the corresponding normalized eigenvector is $\left(1/\sqrt{k}\right)\mathbf{1}_k$.

The operator $\|\cdot\|$ applied to a vector denotes the standard Euclidean 2-norm, while applied to matrices denotes the induced 2-norm, which is equivalent to the matrix spectral radius for symmetric matrices.

We assume that the parameter to be estimated belongs to a subset $\mathcal{U}$ of the Euclidean space $\mathbb{R}^M$. Throughout the paper, the true (but unknown) value of the parameter is denoted by θ^*. We denote a canonical element of $\mathcal{U}$ by θ. The estimate of θ^* at time i at sensor n is denoted by $\mathbf{x}_n(i) \in \mathbb{R}^{M\times 1}$. Without loss of generality, we assume that the initial estimate, $\mathbf{x}_n(0)$, at time 0 at sensor n is a non-random quantity.

Throughout, we assume that all the random objects are defined on a common measurable space, $(\Omega, \mathcal{F})$. In case the true (but unknown) parameter value is θ^*, the probability and expectation operators are denoted by $\mathbb{P}_{\theta^*}[\cdot]$ and $\mathbb{E}_{\theta^*}[\cdot]$, respectively. When the context is clear, we abuse notation by dropping the subscript. Also, all inequalities involving random variables are to be interpreted a.s. (almost surely.)

Spectral graph theory. We review elementary concepts from spectral graph theory. For an *undirected* graph $G = (V, E)$, $V = [1 \cdots N]$ is the set of nodes or vertices, $|V| = N$, and E is the set of edges, $|E| = M$, where $|\cdot|$ is the cardinality. The unordered pair $(n, l) \in E$ if there exists an edge between nodes n and l. We only consider simple graphs, i.e., graphs devoid of self-loops and multiple edges. A graph is connected if there exists a path[5], between each pair of nodes. The neighborhood of node n is

$$\Omega_n = \{l \in V \,|\, (n, l) \in E\} \tag{2}$$

Node n has degree $d_n = |\Omega_n|$ (number of edges with n as one end point.) The structure of the graph is described by the symmetric $N \times N$ adjacency matrix, $A = [A_{nl}]$, $A_{nl} = 1$, if $(n, l) \in E$, $A_{nl} = 0$, otherwise. The degree matrix is the diagonal matrix $D = \text{diag}\,(d_1 \cdots d_N)$. The graph positive semi-definite

[5]A path between nodes n and l of length m is a sequence $(n = i_0, i_1, \cdots, i_m = l)$ of vertices, such that, $(i_k, i_{k+1}) \in E \,\forall\, 0 \le k \le m-1$.

Laplacian matrix, L, and its ordered eigenvalues are

$$L = D - A \tag{3}$$

$$0 = \lambda_1(L) \leq \lambda_2(L) \leq \cdots \leq \lambda_N(L) \tag{4}$$

The smallest eigenvalue $\lambda_1(l)$ is always equal to zero, with $\left(1/\sqrt{N}\right)\mathbf{1}_N$ being the corresponding normalized eigenvector. The multiplicity of the zero eigenvalue equals the number of connected components of the network; for a connected graph, $\lambda_2(L) > 0$. This second eigenvalue is the algebraic connectivity or the Fiedler value of the network; see [17], [18], [19] for detailed treatment of graphs and their spectral theory.

Kronecker product: Since, we are dealing with vector parameters, most of the matrix manipulations will involve Kronecker products. For example, the Kronecker product of the $N \times N$ matrix L and I_M will be an $NM \times NM$ matrix, denoted by $L \otimes I_M$. Denote the $NM \times NM$ matrix $P^{NM} = P_N \otimes I_M = \frac{1}{N}(\mathbf{1}_N \otimes I_M)(\mathbf{1}_N \otimes I_M)^T$. We will deal often with matrices of the form $C = \left[I_{NM} - bL \otimes I_M - aI_{NM} - P^{NM}\right]$, L being a graph Laplacian matrix. It follows from the properties of Kronecker products and the matrices L, P^{NM}, that the eigenvalues of this matrix C are $-a$ and $1 - b\lambda_n(L) - a$, $n \leq i \leq N$, each being repeated M times.

II. Problem Formulation

Let $\theta^* \in \mathbb{R}^{M \times 1}$ be an M-dimensional parameter that is to be estimated by a network of N sensors. We refer to θ as a parameter, although it is a vector of M parameters. Each sensor makes independent observations of noise corrupted linear functions of the parameter. We assume the following observation model for the n-th sensor:

$$\mathbf{z}_n(i) = \overline{H}_n(i)\theta^* + \gamma(i)\zeta_n(i) \tag{5}$$

where: $\left\{\mathbf{z}_n(i) \in \mathbb{R}^{M_n \times 1}\right\}_{i \geq 0}$ is the independent observation sequence for the n-th sensor; $\{\zeta_n(i)\}_{i \geq 0}$ is a zero-mean i.i.d. noise sequence of bounded variance. For most practical sensor network applications, each sensor observes only a subset of M_n of the components of θ, with $M_n \ll M$. Under such a situation, in isolation, each sensor can estimate at most only a part of the parameter. However, if the sensor network is connected in the mean sense (see assumption (A.3)), and under appropriate observability conditions, we will show that it is possible for each sensor to get a consistent estimate of the parameter θ^* by means of local inter-sensor communication.

We formalize the assumptions on global observability, fading signal characteristics and network connectivity:

- **(A.1)Observation Noise**: Recall the observation model in eqn. (5). We assume that the process, $\left\{\zeta(i) = \left[\zeta_1^T(i), \cdots, \zeta_N^T(i)\right]^T\right\}_{i \geq 0}$ is an i.i.d. zero mean process, with finite second moment. The observation noise process, $\{\gamma(i)\zeta(i)\}$, then has non-stationary (in general) characteristics, with variance increasing as $\gamma^2(i)$ over time. The non-decreasing sequence $\{\gamma(i)\}$ models the fading

characteristics of the parameter (signal) over time. In particular, the regime $\gamma(i) \to \infty$ corresponds to the SNR decreasing as $1/\gamma^2(i)$ over time, whereas, $\gamma(i) = 1$ for all i recovers the case of i.i.d. (constant SNR) observations. Also, note that the observation noises at different sensors may be correlated during a particular iteration, we require only temporal independence. The spatial correlation of the observation noise makes our model applicable to practical sensor network problems, for instance, for distributed target localization, where the observation noise is generally correlated across sensors.

The following assumption on the growth rate of $\{\gamma(i)\}$ is imposed throughout:
There exists, $0 \leq \gamma_0 < .5$, such that,

$$\gamma(i) = (i+1)^{\gamma_0}, \quad \forall i \in \mathbb{T}_+ \tag{6}$$

In other words, we assume that the observation noise variance has sublinear growth. The sublinear growth assumption is not restrictive, and as shown in Remark 8 is in fact, necessary for centralized estimators to yield consistent estimates of the parameter.

- **(A.2)Observability:** We require the following global observability condition. The matrix G

$$G = \sum_{n=1}^{N} \overline{H}_n^T \overline{H}_n \tag{7}$$

is full-rank. This distributed observability extends the observability condition for a centralized estimator to get a consistent estimate of the parameter θ^*.

- **(A.3)Random Link Failure:** In digital communications, packets may be lost at random times. To account for this, we let the links (or communication channels among sensors) to fail, so that the edge set and the connectivity graph of the sensor network are time varying. Accordingly, the sensor network at time i is modeled as an undirected graph, $G(i) = (V, E(i))$ and the graph Laplacians as a sequence of i.i.d. Laplacian matrices $\{L(i)\}_{i\geq 0}$. We write

$$L(i) = \overline{L} + \widetilde{L}(i), \ \forall i \geq 0 \tag{8}$$

where the mean $\overline{L} = \mathbb{E}\left[L(i)\right]$. We do not make any distributional assumptions on the link failure model. Although the link failures, and so the Laplacians, are independent at different times, during the same iteration, the link failures can be spatially dependent, i.e., correlated. This is more general and subsumes the erasure network model, where the link failures are independent over space *and* time. Wireless sensor networks motivate this model since interference among the wireless communication channels correlates the link failures over space, while, over time, it is still reasonable to assume that the channels are memoryless or independent.

Connectedness of the graph is an important issue. We do not require that the random instantiations $G(i)$ of the graph be connected; in fact, it is possible to have all these instantiations to be disconnected. We only require that the graph stays connected on *average*. This is captured by

requiring that $\lambda_2\left(\overline{L}\right) > 0$, enabling us to capture a broad class of asynchronous communication models; for example, the random asynchronous gossip protocol analyzed in [20] satisfies $\lambda_2\left(\overline{L}\right) > 0$ and hence falls under this framework.

- **(A.4)Independence Assumptions**: The sequences $\{L(i)\}_{i\in\mathbb{T}_+}$ and $\{\zeta(i)\}_{i\in\mathbb{T}_+}$ are mutually independent.

In Section II-A, we present the algorithm $\mathcal{GLU}$ for distributed parameter estimation with the linear observation model (5). Starting from some initial deterministic estimate of the parameters (the initial states may be random, we assume deterministic for notational simplicity), $\mathbf{x}_n(0) \in \mathbb{R}^{M\times 1}$, each sensor generates by a distributed iterative algorithm a sequence of estimates, $\{\mathbf{x}_n(i)\}_{i\geq 0}$. The parameter estimate $\mathbf{x}_n(i+1)$ at the n-th sensor at time $i+1$ is a function of: its previous estimate; the communicated estimates at time i of its neighboring sensors; and the new observation $\mathbf{z}_n(i)$.

A. Algorithm $\mathcal{GLU}$

Algorithm $\mathcal{GLU}$: Consider the parameter estimation problem with linear observation model (assumptions **(A.1)-(A.2)**). Let $\mathbf{x}(0) = [\mathbf{x}_1(0)^T, \cdots, \mathbf{x}_N(0)^T]^T$ be the initial estimates of θ^* at the sensors. The $\mathcal{GLU}$ algorithm updates the estimate $\mathbf{x}_n(i)$ at sensor n according to the following:

$$\mathbf{x}_n(i+1) = \mathbf{x}_n(i) - \beta(i)\sum_{l\in\Omega_n(i)}\left(\mathbf{x}_n(i) - \mathbf{x}_l(i)\right) + \alpha(i)K\overline{H}_n^T\left(\mathbf{z}_n(i) - \overline{H}_n\mathbf{x}_n(i)\right) \tag{9}$$

The key difference between the above scheme and the $\mathcal{LU}$ in [1] involves the use of different weight sequences for the consensus and the innovation terms, giving the former a mixed time scale behavior. On the other hand, we assume unquantized transmissions in $\mathcal{GLU}$. Another difference is the incorporation of a general matrix gain K into the innovation update. These modifications make the technical analysis of $\mathcal{GLU}$ highly non-trivial and different from that of $\mathcal{LU}$, mostly due to the incorporation of mixed time scale dynamics.

In a compact notation, $\mathcal{GLU}$ may be written as:

$$\mathbf{x}(i+1) = \mathbf{x}(i) - \beta(i)\left(L(i)\otimes I_M\right)\mathbf{x}(i) + \alpha(i)\left(I_N\otimes K\right)\overline{D}_{\overline{H}}\left(\mathbf{z}(i) - D_{\overline{H}}\mathbf{x}(i)\right) \tag{10}$$

We refer to the class of distributed recursive estimation algorithms in (9) as $\mathcal{GLU}$. As will be shown, different choices of the weight sequences $\{\alpha(i)\}, \{\beta(i)\}$ lead to different convergence characteristics of $\mathcal{GLU}$, hence the usage of the term 'class of algorithms'. In the following, we introduce some additional moment requirements and assumptions on the algorithm weight sequences:

- **(A.5)Moment Condition**: There exists $\varepsilon_1 > 0$, such that, the following moment exists:

$$\mathbb{E}_\theta\left[\|\zeta(i)\|^{2+\varepsilon_1}\right] < \infty \tag{11}$$

The above implies the existence of a positive function $\kappa_1(\cdot)$, such that,

$$\mathbb{E}_\theta \left[\left\| \overline{D}_{\overline{H}} \mathbf{z}(i) - \mathbf{1_N} \otimes \left(\left(\frac{1}{N} \mathbf{1}_N \otimes I_M \right) \overline{D}_{\overline{H}} \mathbf{z}(i) \right) \right\|^{2+\varepsilon_1} \right] \leq \gamma^{2+\varepsilon_1}(i) \kappa_1(\theta) < \infty \tag{12}$$

for all $i \in \mathbb{T}_+$. We thus assume the existence of slightly greater than quadratic moment of the observation noise process.

- **(A.6)Weight sequences**: The sequences $\{\alpha(i)\}$ and $\{\beta(i)\}$ are of the form:

$$\alpha(i) = \frac{a}{(i+1)^{\tau_1}}, \qquad \beta(i) = \frac{b}{(i+1)^{\tau_2}} \tag{13}$$

where $a, b > 0$, $0 < \tau_2 \leq \tau_1 \leq 1$. In addition, the weights satisfy the following condition:

$$\tau_1 > \max \left(.5 + \gamma_0, \tau_2 + \gamma_0 + \frac{1}{2+\varepsilon_1} \right) \tag{14}$$

where $\max(\cdot)$ denotes the maximum of $.5 + \gamma_0$ and $\tau_2 + \gamma_0 + \frac{1}{2+\varepsilon_1}$.
The gain matrix K is assumed to be positive definite. To avoid unnecessary technicalities, we also assume that the matrices K and G commute, so that, KG is symmetric positive definite (see [21]). Recall, G to be the invertible Grammian $\sum_{n=1}^{N} \overline{H}_n^T \overline{H}_n$.

Remark 1 We comment on the $\mathcal{GLU}$ assumptions. First, we note that the moment assumption is not restrictive, and most reasonable noise models possess moments of sufficiently high order. Also, it is easy to come up with a choice of algorithm parameters (τ_1, τ_2) given a $0 \leq \gamma < .5$. In fact, any choice of $\tau_1 > .5 + \gamma_0$ suffices, as one can choose τ_2 satisfying $0 < \tau_2 < \tau_1 - .5 - \gamma_0$. That, this choice satisfies assumption (A.6) ((14)), is due to the fact, that, $\frac{1}{2+\varepsilon_1} < .5$ for any $\varepsilon_1 > 0$. Finally, a note on nomenclature. Often, we will use the term $(\tau_1, a, \tau_2, b, K)$-$\mathcal{GLU}$ algorithm to indicate explicitly the $\mathcal{GLU}$ design parameters in force.

Markov. Consider the filtration, $\{\mathcal{F}_i^{\mathbf{x}}\}_{i \geq 0}$, given by

$$\mathcal{F}_i^{\mathbf{x}} = \sigma \left(\mathbf{x}(0), \{L(j), \zeta(j)\}_{0 \leq j < i} \right) \tag{15}$$

It then follows that the random objects $L(i), \mathbf{z}(i)$ are independent of $\mathcal{F}_i^{\mathbf{x}}$, rendering $\{\mathbf{x}(i), \mathcal{F}_i^{\mathbf{x}}\}_{i \in \mathbb{T}_+}$ a Markov process.

B. Centralized linear estimators

The key focus of the paper is to compare the performance achieved by the class of $\mathcal{GLU}$ algorithms to centralized estimation schemes[6]. Specifically, we will restrict this comparison to linear centralized estimators only. To this end, we start by defining *a reasonable* (to be clear soon) class of centralized

[6]A centralized scheme corresponds to a fusion center having access to all sensor observations at all times.

linear[7] of the parameter θ.

Definition 2 (Centralized linear estimator) A centralized linear estimator is a process $\{\mathbf{u}(i)\}_{i\in\mathbb{T}_+}$ evolving as

$$\mathbf{u}(i+1) = \mathbf{u}(i) + \frac{\alpha_c(i)}{N} K_c \sum_{n=1}^{N} \left(\overline{H}_n^T \mathbf{z}_n(i) - \overline{H}_n^T \overline{H}_n \mathbf{u}(i) \right) \tag{16}$$

Here, we assume that the weight sequence $\{\alpha_c(i)\}$ is of the form

$$\alpha_c(i) = \frac{a_c}{(i+1)^{\tau_c}} \tag{17}$$

for some $a_c > 0$ and $\tau_c \geq 0$. Also, K_c is a positive definite gain matrix that commutes with the Grammian G.

A centralized linear estimator is called *good*, if in addition the design parameter satisfies

$$.5 + \gamma_0 < \tau_c \leq 1 \tag{18}$$

Remark 3 We comment on the above definition and justify the nomenclature good. Clearly, different choices of the gain matrix K_c and the weight sequence $\{\alpha_c(i)\}$ would lead to different convergence properties of the estimator $\{\mathbf{u}(i)\}$. As shown in Proposition 7, the condition $.5 + \gamma_0 < \tau_c \leq 1$ is necessary and sufficient for the estimator $\{\mathbf{u}(i)\}$ to be universally[8] consistent from all initial conditions. In particular, the best linear centralized estimator assumes the form in Definition 2 (for a specific choice of K_c and $\{\alpha_c(i)\}$.) Hence, for all purposes, it is sufficient to compare the distributed algorithm $\mathcal{GLU}$ with the class of good centralized estimators defined above. In the following, we will restrict attention to good centralized estimators only, and will often drop the term good when referring to these estimators. Also, similar to the distributed $\mathcal{GLU}$ estimators, we will use the term (τ_c, a_c, K_c) centralized estimator to indicate explicitly the design parameters in force.

Before proceeding to the convergence analysis of $\mathcal{GLU}$ under assumptions **(A.1)-(A.6)**, we establish some properties of general stochastic recursions to be used in the sequel.

III. Some Intermediate Results

We establish three approximation results to be used later. The first one (Lemma 4) is a stochastic analogue of Lemma 18 in [1], the second one (Lemma 5) quantifies the pathwise convergence rate in Lemma 4. Lemma 6 is a time-varying mixed time scale version of Lemma 3 in [1]. Finally, we end this section by listing some convergence properties of the centralized estimators (Definition 2.)

[7]Since we deal with linear centralized estimators only, in the following we drop the term linear when referring to centralized estimators.

[8]By universal consistency of an algorithm, we mean that the algorithm leads to consistent estimates of the parameter θ irrespective of the observation noise distribution, as long as the moment assumption (A.5) is satisfied.

Lemma 4 Consider the scalar time-varying linear system:

$$y(i+1) = (1 - r_1(i))y(i) + r_2(i) \tag{19}$$

Here $\{r_1(i)\}$ is a sequence of independent random variables, such that, $0 \leq r_1(i) \leq 1$ a.s. with mean

$$\overline{r}_1(i) = \frac{a_1}{(i+1)^{\delta_1}} \tag{20}$$

and $a_1 > 0$, $0 \leq \delta_1 \leq 1$. Also, assume $y(0) \geq 0$ and the sequence $\{r_2(i)\}$ is given by

$$r_2(i) = \frac{a_2}{(i+1)^{\delta_2}} \tag{21}$$

where $a_2 > 0, \delta_2 \geq 0$. Then, if $\delta_1 < \delta_2$,

$$\lim_{i \to \infty} y(i) = 0 \text{ a.s.} \tag{22}$$

Proof: The assumptions imply that the sequence $\{y(i)\}$ is non-negative. Define the process $\{V_1(i)\}$ by

$$V_1(i) = y(i) - \sum_{k=0}^{i-1} \left[\left(\prod_{l=k+1}^{i-1} (1 - \overline{r}_1(l)) \right) r_2(k) \right] \tag{23}$$

Since $\delta_1 < \delta_2$, an application of Lemma 18 in [1] yields

$$\lim_{i \to \infty} \sum_{k=0}^{i-1} \left[\left(\prod_{l=k+1}^{i-1} (1 - \overline{r}_1(l)) \right) r_2(k) \right] = 0 \tag{24}$$

Hence, in particular, the second term on the R.H.S. is bounded and $\{y(i)\}$ is well defined. Denote by $\{\mathcal{F}^y(i)\}$ the natural filtration of the process $\{y(i)\}$ and note that $\{V_1(i)\}$ is adapted to this filtration. Using the fact, that

$$\sum_{k=0}^{i} \left[\left(\prod_{l=k+1}^{i} (1 - \overline{r}_1(l)) \right) r_2(k) \right] = (1 - \overline{r}_1(i)) \left[\sum_{k=0}^{i-1} \left[\left(\prod_{l=k+1}^{i-1} (1 - \overline{r}_1(l)) \right) r_2(k) \right] \right] + r_2(i) \tag{25}$$

we have, by the independence condition,

$$\begin{aligned}
\mathbb{E}\left[V_1(i+1) \mid \mathcal{F}^y(i)\right] &= \mathbb{E}\left[y(i+1) \mid \mathcal{F}^y(i)\right] - \sum_{k=0}^{i} \left[\left(\prod_{l=k+1}^{i} (1 - \overline{r}_1(l)) \right) r_2(k) \right] \\
&= (1 - \overline{r}_1(i))y(i) + r_2(i) - \sum_{k=0}^{i} \left[\left(\prod_{l=k+1}^{i} (1 - \overline{r}_1(l)) \right) r_2(k) \right] \\
&= (1 - \overline{r}_1(i))y(i) - \sum_{k=0}^{i-1} \left[\left(\prod_{l=k+1}^{i-1} (1 - \overline{r}_1(l)) \right) r_2(k) \right] \\
&= V_1(i) - \overline{r}_1(i)y(i)
\end{aligned} \tag{26}$$

The nonnegativity of $\{y(i)\}$ implies

$$\mathbb{E}\left[V_1(i+1) \mid \mathcal{F}^y(i)\right] \leq V_1(i) \tag{27}$$

Hence $\{V_1(i)\}$ is a supermartingale. The nonnegativity of $\{y(i)\}$ and the boundedness of the terms $\sum_{k=0}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}(1-\overline{r}_1(l))\right)r_2(k)\right]$ for all i show that $\{V_1(i)\}$ is bounded from below. It then follows that there exists a finite random variable V_1^*, such that,

$$\lim_{i\to\infty} V_1(i) = V_1^* \text{ a.s.} \tag{28}$$

We then have

$$\begin{aligned}\lim_{i\to\infty} y(i) &= \lim_{i\to\infty} V_1(i) + \lim_{i\to\infty}\sum_{k=0}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}(1-\overline{r}_1(l))\right)r_2(k)\right] \\ &= V_1^* \end{aligned} \tag{29}$$

Since $y(0)$ is deterministic, the sequence $\{y(i)\}$ is integrable and we have

$$\mathbb{E}\left[y(i)\right] = \left(\prod_{k=0}^{i}(1-\overline{r}_1(i))\right)y(0) + \sum_{k=0}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}(1-\overline{r}_1(l))\right)r_2(k)\right] \tag{30}$$

An application of Lemma 18 in [1] then shows

$$\lim_{i\to\infty}\mathbb{E}\left[y(i)\right] = 0 \tag{31}$$

and by Fatou's lemma we conclude $\mathbb{E}\left[V_1^*\right] = 0$. Since, V_1^* is nonnegative, being the limit of the nonnegative sequence $\{y(i)\}$, we have

$$V_1^* = 0 \text{ a.s.} \tag{32}$$

and the claim holds. ■

We will also use the following result, which characterizes the convergence rate in the above. The proof is somewhat similar to the arguments in Lemma 4 and we omit it due to space limitations.

Lemma 5 Consider the scalar deterministic time-varying linear system:

$$y(i+1) = (1-r_1(i))y(i) + r_2(i) \tag{33}$$

where the sequences $\{r_1(i)\}$ and $\{r_2(i)\}$ satisfy the hypothesis of Lemma 4.

- **(1)** Then, if $\delta_1 < \delta_2$ and $\delta_1 < 1$,

$$\lim_{i\to\infty}(i+1)^{\delta_0}y(i) = 0 \tag{34}$$

  for all $0 \leq \delta_0 < \delta_2 - \delta_1$.
- **(2)** Let $\delta_1 < \delta_2$ and $\delta_1 = 1$. Then the above conclusion holds, if in addition $a_1 > \delta_0$.
- **(3)** All the above remain valid when $r_1(i)$ is random satisfying the conditions of Lemma 4.

Lemma 6 Under the stated assumptions, there exists i_1 sufficiently large and a constant $c_4 > 0$, such that, for $i \geq i_1$,

$$\mathbf{y}^T \left(\beta(i)\overline{L} \otimes I + \alpha(i)(I_N \otimes K)D_{\overline{H}}\right)\mathbf{y} \geq c_4\alpha(i)\left\|\mathbf{y}\right\|^2, \quad \forall \mathbf{y} \in \mathbb{R}^{NM} \tag{35}$$

Proof: The key difference from the proof of Lemma 3 in [1] is that, the matrix $\left(\beta(i)\overline{L} \otimes I + \alpha(i)(I_N \otimes K)D_{\overline{H}}\right)$ is not symmetric. We first show that the quadratic form

$$\mathbf{y}^T \left(\frac{\beta(i)}{\alpha(i)}\overline{L} \otimes I + (I_N \otimes K)D_{\overline{H}}\right)\mathbf{y} \tag{36}$$

is strictly greater than zero for all $\mathbf{y} \in \mathbb{R}^{NM}$ satisfying $\|\mathbf{y}\| = 1$ for all sufficiently large i. To this end, for such $\mathbf{y}$, consider the decomposition

$$\mathbf{y} = \mathbf{y}_{\mathcal{C}} + \mathbf{y}_{\mathcal{C}^\perp} \tag{37}$$

Define the symmetric matrix D_K by

$$D_K = \frac{1}{2}\left[(I_N \otimes K)D_{\overline{H}}\right] + \frac{1}{2}\left[(I_N \otimes K)D_{\overline{H}}\right]^T \tag{38}$$

Noting that

$$\mathbf{y}^T\left[(I_N \otimes K)D_{\overline{H}}\right]\mathbf{y} = \mathbf{y}^T D_K \mathbf{y} \tag{39}$$

we have

$$\begin{aligned}
\mathbf{y}^T \left(\frac{\beta(i)}{\alpha(i)}\overline{L} \otimes I + (I_N \otimes K)D_{\overline{H}}\right)\mathbf{y} &= \mathbf{y}^T \left(\frac{\beta(i)}{\alpha(i)}\overline{L} \otimes I + D_K\right)\mathbf{y} \\
&= \mathbf{y}^T \left(\frac{\beta(i)}{\alpha(i)}\overline{L} \otimes I\right)\mathbf{y} + \mathbf{y}^T D_K \mathbf{y} \\
&= \mathbf{y}_{\mathcal{C}^\perp}^T \left(\frac{\beta(i)}{\alpha(i)}\overline{L} \otimes I\right)\mathbf{y}_{\mathcal{C}^\perp} + \mathbf{y}_{\mathcal{C}^\perp}^T D_K \mathbf{y}_{\mathcal{C}^\perp} \\
&\quad + 2\mathbf{y}_{\mathcal{C}^\perp}^T D_K \mathbf{y}_{\mathcal{C}} + \mathbf{y}_{\mathcal{C}}^T D_K \mathbf{y}_{\mathcal{C}} \\
&\geq \frac{\beta(i)}{\alpha(i)}\lambda_2(\overline{L})\left\|\mathbf{y}_{\mathcal{C}^\perp}\right\|^2 + \mathbf{y}_{\mathcal{C}^\perp}^T D_K \mathbf{y}_{\mathcal{C}^\perp} \\
&\quad + 2\mathbf{y}_{\mathcal{C}^\perp}^T D_K \mathbf{y}_{\mathcal{C}} + \mathbf{y}_{\mathcal{C}}^T D_K \mathbf{y}_{\mathcal{C}}
\end{aligned} \tag{40}$$

Now, the symmetricity of D_K implies the existence of a constant $c_{15} > 0$, large enough, such that,

$$\mathbf{y}_{\mathcal{C}^\perp}^T D_K \mathbf{y}_{\mathcal{C}^\perp} \geq -c_{15}\left\|\mathbf{y}_{\mathcal{C}^\perp}\right\|^2 \tag{41}$$

$$\mathbf{y}_{\mathcal{C}^\perp}^T D_K \mathbf{y}_{\mathcal{C}} \geq -c_{15}\left\|\mathbf{y}_{\mathcal{C}}\right\|\left\|\mathbf{y}_{\mathcal{C}^\perp}\right\| \tag{42}$$

Also, using the form $\mathbf{y}_{\mathcal{C}} = \mathbf{1}_N \otimes \mathbf{a}$, for some $\mathbf{a} \in \mathbb{R}^M$, we note that

$$\begin{aligned}
\mathbf{y}_{\mathcal{C}}^T D_K \mathbf{y}_{\mathcal{C}} &= \mathbf{y}_{\mathcal{C}}^T \left[(I_N \otimes K) D_{\overline{H}}\right] \mathbf{y}_{\mathcal{C}} \\
&= \sum_{n=1}^{N} \mathbf{a}^T K \overline{H}_n \mathbf{a} \\
&= \mathbf{a} K G \mathbf{a} \\
&\geq \lambda_{\min} \|\mathbf{a}\|^2 \\
&= \frac{\lambda_{\min}}{N} \|\mathbf{y}_{\mathcal{C}}\|^2
\end{aligned} \tag{43}$$

where the last but one step uses the fact, that the matrix KG is positive definite, as both K and G are positive definite and they commute. Note, in particular, that $\lambda_{\min} > 0$. Substituting the above in eqn. (40), we have

$$\mathbf{y}^T \left(\frac{\beta(i)}{\alpha(i)} \overline{L} \otimes I + (I_N \otimes K) D_{\overline{H}} \right) \mathbf{y} \geq \left(\frac{\beta(i)}{\alpha(i)} \lambda_2(\overline{L}) - c_{15} \right) \|\mathbf{y}_{\mathcal{C}^\perp}\|^2 - 2c_{15} \|\mathbf{y}_{\mathcal{C}}\| \|\mathbf{y}_{\mathcal{C}^\perp}\| + \frac{\lambda_{\min}}{N} \|\mathbf{y}_{\mathcal{C}}\|^2 \tag{44}$$

Since $\lim_{i \to \infty} \beta(i)/\alpha(i) = \infty$ ($\tau_2 < \tau_1$), we can choose i_1 large enough, such that, for $i \geq i_0$

$$\frac{\beta(i)}{\alpha(i)} \lambda_2(\overline{L}) - c_{15} > 0 \tag{45}$$

$$\frac{\lambda_{\min}}{N} \left[\frac{\beta(i)}{\alpha(i)} - c_{15} \right] > c_{15}^2 \tag{46}$$

We now verify the claim in eqn. (36) for $i \geq i_1$. Clearly, if $\mathbf{y}_{\mathcal{C}} = \mathbf{0}$, the quadratic form reduces to

$$\mathbf{y}^T \left(\frac{\beta(i)}{\alpha(i)} \overline{L} \otimes I + (I_N \otimes K) D_{\overline{H}} \right) \mathbf{y} \geq \left(\frac{\beta(i)}{\alpha(i)} \lambda_2(\overline{L}) - c_{15} \right) \|\mathbf{y}_{\mathcal{C}^\perp}\|^2 = \frac{\beta(i)}{\alpha(i)} \lambda_2(\overline{L}) - c_{15} > 0 \tag{47}$$

(Note that, the constraint that $\mathbf{y}$ lies on the unit circle forces $\|\mathbf{y}_{\mathcal{C}^\perp}\|$ to be 1, if $\mathbf{y}_{\mathcal{C}} = \mathbf{0}$.) On the other hand, if $\mathbf{y}_{\mathcal{C}} > 0$, we have

$$\mathbf{y}^T \left(\frac{\beta(i)}{\alpha(i)} \overline{L} \otimes I + (I_N \otimes K) D_{\overline{H}} \right) \mathbf{y} \geq \|\mathbf{y}_{\mathcal{C}}\|^2 \left[\left(\frac{\beta(i)}{\alpha(i)} \lambda_2(\overline{L}) - c_{15} \right) \frac{\|\mathbf{y}_{\mathcal{C}^\perp}\|^2}{\|\mathbf{y}_{\mathcal{C}}\|^2} - 2c_{15} \frac{\|\mathbf{y}_{\mathcal{C}^\perp}\|}{\|\mathbf{y}_{\mathcal{C}}\|} + \frac{\lambda_{\min}}{N} \right] \tag{48}$$

The term on the R.H.S. is always strictly greater than zero by the discriminant condition of eqn. (45).

The assertion in eqn. (36) thus holds. Since the quadratic form is a continuous function of $\mathbf{y}$, its positivity on the unit circle implies, there exists $c_4 > 0$, such that,

$$\inf_{\|\mathbf{y}\|=1} \mathbf{y}^T \left(\frac{\beta(i)}{\alpha(i)} \overline{L} \otimes I + (I_N \otimes K) D_{\overline{H}} \right) \mathbf{y} \geq c_4 > 0 \tag{49}$$

It then follows that, for all $\mathbf{y} \in \mathbb{R}^{NM}$,

$$\mathbf{y}^T \left(\frac{\beta(i)}{\alpha(i)} \overline{L} \otimes I + (I_N \otimes K) D_{\overline{H}} \right) \mathbf{y} \geq c_4 \|\mathbf{y}\|^2 \tag{50}$$

and hence

$$\begin{aligned}\mathbf{y}^T\left(\beta(i)\overline{L}\otimes I+\alpha(i)(I_N\otimes K)D_{\overline{H}}\right)\mathbf{y} &= \alpha(i)\mathbf{y}^T\left(\frac{\beta(i)}{\alpha(i)}\overline{L}\otimes I+(I_N\otimes K)D_{\overline{H}}\right)\mathbf{y}\\ &\geq \alpha(i)c_4\left\|\mathbf{y}\right\|^2 \qquad (51)\end{aligned}$$

for $i \geq i_1$. ■

Note that, the condition $\lim_{i\to\infty}\beta(i)/\alpha(i)=\infty$ is required for Lemma 6.

The following proposition justifies the nomenclature good in Definition 2. In particular, it shows that under assumptions (A.1),(A.2),(A.5), there exists a noise distribution (Gaussian), such that, the centralized scheme is not consistent if τ_c fails to satisfy the requirement (18).

Proposition 7 (1) Suppose the process $\{\zeta(i)\}$ is Gaussian. Consider the centralized estimator $\{\mathbf{u}(i)\}$. Then, if $\tau_c \leq \gamma_0 + .5$ or $\tau_c > 1$, the sequence $\{\mathbf{u}(i)\}$ is not consistent from arbitrary initial condition $\mathbf{u}(0)$.

(2) Let assumptions (A.1),(A.2),(A.5) hold. Then, a good centralized estimator is consistent (universally) from all initial conditions.

(3) Let assumptions (A.1),(A.2),(A.5) hold. Consider a good centralized estimator with design parameters (τ_c, a_c, K_c). Then, there exists a $(\tau_1, a, \tau_2, b, K)$-$\mathcal{GLU}$ estimator, such that, $\tau_1 = \tau_c$, $a = a_c$, $K = K_c$.

Remark 8 As a consequence of the first assertion, we note that, for a centralized linear estimator to achieve consistency, the parameter γ_0 should be strictly less than .5.

Proof: Due to space limitations, we omit the proof which follows from standard properties of stochastic recurrences and approximation ([22]).

We present an intuitive sketch of the proof of the first assertion. From (16), we note that, at time i, an observation noise is incorporated on the right hand side (R.H.S.) with variance of the order $(i+1)^{2\gamma_0-2ta u_c}$. Clearly, if $\tau_c \leq .5+\gamma_0$, as $i\to\infty$ the cumulative noise adds up to ∞. For Gaussian noise, this would lead to unboundedness of the estimate sequence $\{\mathbf{u}(i)\}$. This explains the lower bound in the choice of τ_c. On the other hand, if $\tau_c > 1$, the $\{\alpha_c\}$ becomes summable and the updates die out quickly. Hence, depending on the initial estimate $\mathbf{u}(0)$, it may not be possible to progress towards θ^*. Thus, in general, we need $\tau_c \leq 1$.

The second assertion follows from standard stochastic approximation arguments (see, for example [22] and Theorem 1 in [23].)

The third assertion simply states that there exists a choice of τ_2 satisfying assumption (A.6), when $\tau_1 = \tau_c$ and $K = K_c$. This is immediate from Remark 1. ■

In the case $\gamma_0 = 1$, i.e., the observation process is stationary (constant SNR), the following property of $\{\mathbf{u}(i)\}$ holds:

Proposition 9 Suppose $\gamma_0 = 0$ and assumptions (A.1),(A.2),(A.5) hold. Then, in addition to the consistency in Proposition 7, we have the following:

(1) Assume $\tau_c = 1$, i.e., the weight sequence $\{\alpha_c(i)\}$ is of the form

$$\alpha_c(i) = \frac{a_c}{i+1} \tag{52}$$

Then, if $a_c > \frac{N}{2\lambda_{\min}(KG)}$, the normalized sequence $\{1/\sqrt{(i+1)}(\mathbf{u}(i) - \theta^*)\}$ is asymptotically normal, i.e.,

$$\sqrt{(i+1)}\,(\mathbf{u}(i) - \theta^*) \Longrightarrow \mathcal{N}(\mathbf{0}, S_c(K)) \tag{53}$$

where, the asymptotic variance is given by:

$$S_c(K) = \frac{a^2}{N^2}\int_0^\infty e^{\Sigma_1 v} S_1 e^{\Sigma^T v} dv \tag{54}$$

$$\Sigma_1 = -\frac{a}{N}KG + \frac{1}{2}I_M \tag{55}$$

$$S_1 = K\,(\mathbf{1}_N \otimes I_M)^T\,\overline{D}_{\overline{H}} S_\zeta \overline{D}_{\overline{H}}^T\,(\mathbf{1}_N \otimes I_M)\,K^T \tag{56}$$

(2) Let the hypothesis of the previous assertion hold and choose $K_c = K_c^* = G^{-1}$. Then, the estimator $\{\mathbf{u}(i)\}$ is the best linear centralized estimator in terms of asymptotic variance irrespective of the distribution of the observation noise $\zeta(i)$. In addition, if the observation noise sequence $\{\zeta(i)\}$ is Gaussian, $\{\mathbf{u}(i)\}$ as defined above, is the optimum centralized estimator, whose asymptotic variance $S_c(K^*)$ equals the centralized Fisher information rate.

Proof: The proof of the first assertion is omitted due to space limitations (see [1] for similar arguments.) That, $K_c = G^{-1}$ yields the best linear estimator is standard (see, for example, [24].) ■

IV. Main Results

Theorem 10 Consider a fixed $0 \le \gamma_0 < .5$. Let assumptions (A.1),(A.2),(A.5) hold.

(1) Consider the $\mathcal{GLU}$ algorithm with design parameters $(\tau_1, a, \tau_2, b, K)$ satisfying assumption (A.6). For each sensor n, the estimate sequence $\{\mathbf{x}_n(i)\}$ generated by the $\mathcal{GLU}$ is a consistent estimator of θ^*, i.e.,

$$\mathbb{P}_{\theta^*}\left(\lim_{i\to\infty} \mathbf{x}_n(i) = \theta^*\right) = 1, \quad \forall n \tag{57}$$

(2) Consider a centralized estimator $\{\mathbf{u}(i)\}$ corresponding to a given choice of $\{\alpha_c\}$ and K_c. Choose $K = K_c$, $\tau_1 = \tau_c$ and τ_2 satisfying $0 < \tau_2 < \tau_1 - \gamma_0 - \frac{1}{2+\varepsilon_1}$, such that, assumptions (A.1)-(A.6) hold (such a choice is always possible by Proposition 7.) Also, if $\tau_1 = 1$, further assume that the constant a in assumption **(A.6)** satisfies

$$a > \frac{N\tau_0}{\lambda_{\min}(KG)} \tag{58}$$

For each sensor n, consider the estimate sequence $\{\mathbf{x}_n(i)\}$ generated by the corresponding $\mathcal{GLU}$ algorithm with the above design parameters. Then, for every $0 \leq \tau_0 < \tau_1 - \tau_2 - \frac{1}{2+\varepsilon_1}$, we have

$$\mathbb{P}_{\theta^*}\left(\lim_{i\to\infty}(i+1)^{\tau_0}\left(\mathbf{x}_n(i)-\mathbf{u}(i)\right)=0\right)=1, \quad \forall n \tag{59}$$

We discuss the consequences of Theorem 10. The first assertion states that, as long as $0 \leq \gamma_0 < .5$, any distributed $\mathcal{GLU}$ estimator yields consistent parameter estimates at every sensor. By Remark 8, this is precisely the class of fading parameters, a centralized estimator can estimate consistently. In other words, as long as a centralized linear estimator can consistently estimate a parameter, a distributed $\mathcal{GLU}$ estimator can. This is interesting, as the range of allowable γ_0s is independent of the network topology, and any random network satisfying the mean connectivity is sufficient. The second assertion quantifies the rate at which the distributed $\mathcal{GLU}$ estimator converges to the centralized estimator. Again, this rate is independent of the network topology.

The following result (Theorem 11) shows in what sense the $\mathcal{GLU}$ algorithm is optimal. We assume $\gamma_0 = 0$ in what follows. Suitable extensions to arbitrary γ_0 may be possible, however, this would impose added technicalities and digress from the main focus of the paper. Also, the notion of asymptotic variance as the metric for comparing different consistent estimators, is not quite clear for nonstationary recursive procedures.

Theorem 11 (1) Recall the positive definite matrix $G = \sum_{n=1}^{N} \overline{H}_n^T \overline{H}_n$. Assume $\tau_1 = 1$, i.e., the weight sequence $\{\alpha(i)\}$ is of the form

$$\alpha(i) = \frac{a}{i+1} \tag{60}$$

where $a > \frac{N}{2\lambda_{\min}(KG)}$ and K is the positive definite matrix gain that commutes with G. Choose any τ_2 satisfying

$$\tau_2 + \frac{1}{2+\varepsilon_1} < .5 \tag{61}$$

and note that such a choice exists as $\frac{1}{2+\varepsilon_1} < .5$. Consider the $\mathcal{GLU}$ algorithm with design parameters $(\tau_1, a, \tau_2, b, K)$ chosen above (this ensures that $(\tau_1, a, \tau_2, b, K)$ satisfy assumption (A.6).) Then, the normalized estimate sequence $\{1/\sqrt{(}i+1)(\mathbf{x}_n(i)-\theta^*)\}$ is asymptotically normal for each n, i.e.,

$$\sqrt{(}i+1)\left(\mathbf{x}_n(i)-\theta^*\right) \Longrightarrow \mathcal{N}(\mathbf{0}, S_c(K)) \tag{62}$$

Here, the asymptotic variance $S_c(K)$ is the same obtained by a centralized estimator in Theorem 9 with gain $K_c = K$.

(2) Let the hypothesis of the previous assertion hold with the matrix gain K taking the value $K^* = G^{-1}$. Then, the asymptotic variance at each sensor is $S_c(K^*)$, which is the asymptotic variance achieved by the best linear centralized estimator (see Proposition 9.) In particular, if the observation noise process is Gaussian, the $\mathcal{GLU}$ estimator constructed above is asymptotically efficient.

We interpret the above. The first assertion implies that given a centralized estimator with matrix gain K and satisfying the assumptions in Proposition 9, there exists a distributed $\mathcal{GLU}$ estimator achieving the same asymptotic variance $S_c(K)$. This result is remarkable, as the asymptotic variance $S_c(K)$ is independent of the network topology $\overline{L}$. This is possible due to the mixed time scale behavior resulting from appropriate choice of τ_1, τ_2. This invariance to the network topology is not achievable by the single time scale scheme ($\tau_1 = \tau_2$) developed in [1]. In a sense, Theorem 11 justifies the applicability and advantage of distributed estimation schemes. Apart from issues of robustness, implementing a centralized estimator is much more communication intensive as it requires transmitting all sensor data to a fusion center at all times. On the other hand, the distributed $\mathcal{GLU}$ algorithm requires only sparse local communication among the sensors at each step, and achieves the performance of a centralized estimator asymptotically. The second assertion of the theorem reemphasizes the optimality and applicability of distributed estimation schemes, and shows that $\mathcal{GLU}$ can be designed to achieve the asymptotic variance of the optimal linear centralized scheme. In particular, if the observation noise process is Gaussian, $\mathcal{GLU}$ leads to asymptotically efficient estimators at each sensor.

V. $\mathcal{GLU}$: Convergence properties

As noted earlier, the mixed time scale behavior of $\mathcal{GLU}$ does not permit the use of standard stochastic approximation tools for establishing convergence. Moreover, to be able to establish important qualitative properties like asymptotic time scale separation, we need to clearly distinguish the long term effects of the consensus and innovations potential. We briefly outline the key steps involved in such a pursuit. We first identify conditions under which the sensor estimates $\{\mathbf{x}_n(i)\}$ converge to an *averaged* estimate $\{\mathbf{x}_{\text{avg}}(i)\}$ over the network and recognize the pathwise (strong) convergence rate. This is carried out in Lemma 15. The averaged estimator $\{\mathbf{x}_{\text{avg}}(i)\}$ is not quite the centralized estimator $\{\mathbf{u}(i)\}$, the key reason being the averaged local innovations is not the centralized innovation. This leads us to study the rate of convergence of the averaged local innovations to the centralized innovation and hence, the convergence rate of the averaged estimate sequence to the centralized. This is accomplished in Lemma 16. The analysis in all these steps culminate to Theorems 10,11, the main results of the paper. These results identify conditions under which the consistent estimate sequences $\{\mathbf{x}_n(i)\}$ inherit the centralized convergence rate to θ^*. In particular, they establish sufficient conditions for the equivalence between the distributed and centralized schemes in terms of asymptotic variance. The methodology developed in this work is of independent interest and goes beyond the setting of distributed parameter estimation. We envision its applicability in the analysis of generic dynamical systems interacting over a network.

In what follows, we consider the $\mathcal{GLU}$ algorithm with fixed design parameters $(\tau_1, a, \tau_2, b, K)$ and assumptions (A.1)-(A.6) hold throughout.

We start by establishing pathwise boundedness of the sequence $\{\mathbf{x}(i)\}$.

Lemma 12 There exists a finite random variable $R > 0$, such that,

$$\mathbb{P}_{\theta^*}\left(\sup_{i\in\mathbb{T}_+}\|\mathbf{x}(i)\| \leq R\right) = 1 \tag{63}$$

Proof: Define the process $\{\mathbf{y}(i)\}$ as

$$\mathbf{y}(i) = \mathbf{x}(i) - \mathbf{1}_N \otimes \theta^* \tag{64}$$

The assertion would follow if we establish boundedness for the process $\{\mathbf{y}(i)\}$. From eqn. (10) we note that $\{y(i)\}$ satisfies the recursion:

$$\begin{aligned}\mathbf{y}(i+1) &= \left(I_{NM} - \beta(i)\overline{L}\otimes I_M - \alpha(i)(I_N\otimes K)D_{\overline{H}}\right)\mathbf{y}(i) - \beta(i)\left(\widetilde{L}(i)\otimes I_M\right)\mathbf{y}(i)\\ &\quad +\alpha(i)(I_N\otimes K)\left(\overline{D}_{\overline{H}}\mathbf{z}(i) - D_{\overline{H}}(\mathbf{1}_N\otimes\theta^*)\right)\end{aligned} \tag{65}$$

where we use the invariance of the Laplacian operator,

$$\left(\overline{L}\otimes I_M\right)(\mathbf{1}_N\otimes\theta^*) = \mathbf{0}_{NM}$$

Consider the process $\{V_2(i)\}$ given by

$$V_2(i) = \|\mathbf{y}(i)\|^2 \tag{66}$$

By using the conditional independence properties, it can be shown that,

$$\begin{aligned}\mathbb{E}_{\theta^*}\left[V_2(i+1)\mid\mathcal{F}_i\right] &= V(i) + \beta^2(i)\mathbf{y}(i)^T\mathbb{E}_{\theta^*}\left[\widetilde{L}^2(i)\right]\mathbf{y}(i) + \alpha^2(i)\mathbb{E}_{\theta^*}\left[\left\|\overline{D}_{\overline{H}}\mathbf{z}(i) - D_{\overline{H}}(\mathbf{1}_N\otimes\theta^*)\right\|^2\right]\\ &\quad -2\mathbf{y}^T(i)\left(\beta(i)\overline{L}\otimes I_M + \alpha(i)\left(I_N\otimes K\right)D_{\overline{H}}\right)\mathbf{y}(i) + \beta^i\mathbf{y}^T(i)(\overline{L}\otimes I_M)^2\mathbf{y}(i)\\ &\quad +\alpha^2(i)\mathbf{y}^T(i)\left(\left(I_N\otimes K\right)D_{\overline{H}}\right)^T\left(\left(I_N\otimes K\right)D_{\overline{H}}\right)\mathbf{y}(i)\\ &\quad +2\alpha(i)\beta(i)\mathbf{y}^T(i)(\overline{L}\otimes I_M)(I_N\otimes K)\mathbf{y}(i)\end{aligned} \tag{67}$$

We use the following inequalities:

$$\begin{aligned}\mathbf{y}(i)^T\mathbb{E}_{\theta^*}\left[\widetilde{L}^2(i)\right]\mathbf{y}(i) &= \mathbf{y}_{\mathcal{C}^\perp}^T(i)\mathbb{E}_{\theta^*}\left[\widetilde{L}^2(i)\right]\mathbf{y}_{\mathcal{C}^\perp}(i)\\ &\leq c_5\left\|\mathbf{y}_{\mathcal{C}^\perp}(i)\right\|^2 \qquad (68)\\ \mathbf{y}^T(i)(\overline{L}\otimes I_M)^2\mathbf{y}(i) &= \mathbf{y}_{\mathcal{C}^\perp}^T(i)(\overline{L}\otimes I_M)^2\mathbf{y}_{\mathcal{C}^\perp}(i)\\ &\leq \lambda_N^2(\overline{L})\left\|\mathbf{y}_{\mathcal{C}^\perp}(i)\right\|^2 \qquad (69)\\ 2\mathbf{y}^T(i)\left(\beta(i)\overline{L}\otimes I_M + \alpha(i)\left(I_N\otimes K\right)D_{\overline{H}}\right)\mathbf{y}(i) &\geq \beta(i)\mathbf{y}^T(i)(\overline{L}\otimes I_M)\mathbf{y}(i) + \mathbf{y}^T(i)\left(\beta(i)\overline{L}\otimes I_M\right.\\ &\quad \left.+\alpha(i)\left(I_N\otimes K\right)D_{\overline{H}}\right)\mathbf{y}(i)\\ &\geq \beta(i)\lambda_2(\overline{L})\left\|\mathbf{y}_{\mathcal{C}^\perp}(i)\right\|^2 + c_4\alpha(i)\left\|\mathbf{y}(i)\right\|^2 \qquad (70)\end{aligned}$$

We use Lemma 6 to obtain the last inequality. Introducing additional constants to bound the quadratic

forms and the moments, we derive the following from eqn. (67):

$$\begin{aligned}\mathbb{E}_{\theta^*}\left[V_2(i+1) \mid \mathcal{F}_i\right] &\leq V_2(i) - \left(\beta(i)\lambda_2(\overline{L}) - \beta^2(i)c_5 - \beta^2(i)\lambda_N^2(\overline{L})\right)\left\|\mathbf{y}_{\mathcal{C}^\perp}(i)\right\|^2 \\ &\quad - \left(c_4\alpha(i) - \alpha(i)\beta(i)c_7\right)\left\|\mathbf{y}(i)\right\|^2 + \alpha^2(i)\gamma^2(i)c_8 + \alpha^2(i)c_6\left\|\mathbf{y}(i)\right\|^2 \quad (71)\end{aligned}$$

where $c_8 > 0$ is a constant, such that,

$$\alpha^2(i)\mathbb{E}_{\theta^*}\left[\left\|\overline{D}_{\overline{H}}\mathbf{z}(i) - D_{\overline{H}}(\mathbf{1}_N \otimes \theta^*)\right\|^2\right] = \alpha^2(i)\gamma^2(i)c_8 \quad (72)$$

Since $\beta^2(i)$ goes to zero faster than $\beta(i)$, the $\beta(i)$ term dominates in the second expression of eqn. (71) eventually. Similarly, the $\alpha(i)$ term dominates the third expression eventually. Choose $c_9 = \max(c_6, c_8)$. Since, $\gamma(i) \geq 1$ (assumption (A.2)), there exists i_2 large enough, such that, for $i \geq i_2$

$$\begin{aligned}\mathbb{E}_{\theta^*}\left[V_2(i+1) \mid \mathcal{F}_i\right] - V_2(i) &\leq \alpha^2(i)\gamma^2(i)c_8 + \alpha^2(i)c_6V_2(i) \\ &\leq c_9\alpha^2(i)\gamma^2(i)(1+V_2(i)) \quad (73)\end{aligned}$$

Now introduce the process

$$\widetilde{V}_2(i) = (1+V_2(i))\prod_{k=i}^{\infty}(1+c_9\alpha^2(k)\gamma^2(k)) \quad (74)$$

Note that the above is well defined as the product $\prod_{k=i}^{\infty}(1+c_9\alpha^2(k)\gamma^2(k))$ converges for all i due to the square summability of $\{\alpha(i)\gamma(i)\}$ (assumption (A.6)). Eqn. (73) and some algebraic manipulations lead to

$$\mathbb{E}_{\theta^*}\left[\widetilde{V}_2(i+1) \mid \mathcal{F}_i\right] \leq \widetilde{V}_2(i) \quad (75)$$

thus establishing that the sequence $\{\widetilde{V}_2(i)\}$ is a nonnegative supermartingale. Hence, there exists a finite random variable $\widetilde{R}$, such that, $\lim_{i\to\infty}\widetilde{V}_2(i) = R$ a.s. We then have from eqn. (74)

$$\lim_{i\to\infty} V_2(i) = \widetilde{R} - 1 \text{ a.s.} \quad (76)$$

Hence, $\{V_2(i)\}$ is bonded pathwise and the assertion follows. ■

Remark 13 A deeper investigation of the supermartingale would reveal that $V_2(i)$ in fact, converges to zero. This would have established the consistency of the estimators. However, to obtain strong convergence rates, we need to study the sample paths more critically. The rest of this subsection is devoted to this study.

The following lemma identifies the rate at which the estimates converge to a network *averaged estimate* and hence characterizes the information flow in the network.

Before that, we establish the following:

Proposition 14 Let assumptions (A.1)-(A.6) hold.

(1) For all $i \in \mathbb{T}_+$, define

$$J_1(\mathbf{z}(i)) = (I_N \otimes K)\overline{D}_{\overline{H}}\mathbf{z}(i) - \mathbf{1_N} \otimes \left(\left(\frac{1}{N}\mathbf{1}_N \otimes I_M\right)(I_N \otimes K)\overline{D}_{\overline{H}}\mathbf{z}(i)\right) \tag{77}$$

Then, we have the following:

$$\mathbb{P}_{\theta^*}\left(\frac{1}{(i+1)^{\gamma_0+\frac{1}{2+\varepsilon_1}+\delta}}\left\|J_1(\mathbf{z}(i))\right\| = 0\right) = 0 \tag{78}$$

(2) Recall the matrix,

$$P^{NM} = \frac{1}{N}\left(\mathbf{1}_N \otimes I_M\right)\left(\mathbf{1}_N \otimes I_M\right)^T \tag{79}$$

Then, for $i \in \mathbb{T}_+$ sufficiently large, we have

$$\left\|I_{NM} - \beta(i)\left(L(i) \otimes I_M\right) - P^{NM}\right\| = 1 - \beta(i)\lambda_2(L(i)) \tag{80}$$

Proof: For the first assertion, consider any $\varepsilon_2 > 0$. By Chebyshev's inequality and assumption (A.5),

$$\begin{aligned}\mathbb{P}_{\theta^*}\left(\frac{1}{(i+1)^{\frac{1}{2+\varepsilon_1}+\delta}}\left\|J_1(\mathbf{z}(i))\right\| > \varepsilon_2\right) &\leq \frac{1}{\varepsilon_2^{2+\varepsilon_1}(i+1)^{1+(\delta+\gamma_0)(2+\varepsilon_1)}}\mathbb{E}_{\theta^*}\left[\left\|J_1(\mathbf{z}(i))\right\|^{2+\varepsilon_1}\right] \\ &= \frac{\kappa(\theta^*)}{\varepsilon_2^{2+\varepsilon_1}}\frac{1}{(i+1)^{1+\delta(2+\varepsilon_1)}}\end{aligned} \tag{81}$$

Since, $\delta > 0$, the sequence $\{\frac{1}{(i+1)^{1+\delta(2+\varepsilon_1)}}\}$ is square summable and we obtain

$$\sum_{i\in\mathbb{T}_+}\mathbb{P}_{\theta^*}\left(\frac{1}{(i+1)^{\frac{1}{2+\varepsilon_1}+\delta}}\left\|J_1(\mathbf{z}(i))\right\| > \varepsilon_2\right) < \infty \tag{82}$$

It then follows from the Borel-Cantelli lemma (see [25]) that,

$$\mathbb{P}_{\theta^*}\left(\frac{1}{(i+1)^{\frac{1}{2+\varepsilon_1}+\delta}}\left\|J_1(\mathbf{z}(i))\right\| > \varepsilon_2 \text{ i.o.}\right) = 0 \tag{83}$$

where i.o. stands for infinitely often. Since the above holds for $\varepsilon_2 > 0$ arbitrarily small, the claim in eqn. (78) holds by standard arguments.

For the second assertion, we note from the discussion on Kronecker products in Section I-B that, the eigenvalues of the matrix $\left(I_{NM} - \beta(i)\left(L(i) \otimes I_M\right) - P^{NM}\right)$ are 0 and $1-\beta(i)\lambda_n(L(i))$, $i = 2, \cdots, N$, each repeated M times. Since, the Laplacian eigenvalues are all bounded above by N^2 and $\beta(i) \to 0$, there exists $i_4 \in \mathbb{T}_+$ sufficiently large, such that, for $i \geq i_4$, $\beta(i)\lambda_n(L(i)) < 1$, for all $2 \leq n \leq N$. The assertion is then obvious. ■

Lemma 15 Define the averaged estimate sequence $\{\mathbf{x}_{\text{avg}}(i)\}$ as

$$\mathbf{x}_{\text{avg}}(i) = \frac{1}{N}(\mathbf{1}_N \otimes I_M)\mathbf{x}(i) \tag{84}$$

Then for every τ_0, such that,

$$0 \leq \tau_0 < \tau_1 - \tau_2 - \gamma_0 - \frac{1}{2+\varepsilon} \tag{85}$$

we have

$$\mathbb{P}_{\theta^*}\left(\lim_{i\to\infty}(i+1)^{\tau_0}\left(\mathbf{x}(i) - \mathbf{1}_N \otimes \mathbf{x}_{\text{avg}}(i)\right) = 0\right) = 1 \tag{86}$$

Proof: Define the process $\{\mathbf{y}_1(i)\}$:

$$\widehat{\mathbf{y}}(i) = \mathbf{x}(i) - \mathbf{1}_N \otimes \mathbf{x}_{\text{avg}}(i) \tag{87}$$

Recall the matrix

$$P^{NM} = \frac{1}{N}\left(\mathbf{1}_N \otimes I_M\right)\left(\mathbf{1}_N \otimes I_M\right)^T \tag{88}$$

and note that

$$P^{NM}\mathbf{x}(i) = \mathbf{1}_N \otimes \mathbf{x}_{\text{avg}}(i), \quad P^{NM}\left(\mathbf{1}_N \otimes \mathbf{x}_{\text{avg}}(i)\right) = \mathbf{1}_N \otimes \mathbf{x}_{\text{avg}}(i) \tag{89}$$

From eqn. (10) we then note that $\{\mathbf{y}_1(i)\}$ satisfies the recursion:

$$\begin{aligned}\widehat{\mathbf{y}}(i+1) \;=\; & \left(I_{NM} - \beta(i)\overline{L}\otimes I_M - P^{NM}\right)\widehat{\mathbf{y}}(i) - \alpha(i)\left[(I_N\otimes K)D_{\overline{H}}\mathbf{x}(i)\right. \\ & \left. -\mathbf{1}_N \otimes \left(\frac{1}{N}(\mathbf{1}_N\otimes I_M)(I_N\otimes K)D_{\overline{H}}\mathbf{x}(i)\right)\right] \\ & +\alpha(i)\left[J_1(\mathbf{z}(i))\right]\end{aligned} \tag{90}$$

where $J_1(\mathbf{z}(i))$ is defined in (77). Choose δ satisfying

$$0 < \delta < \tau_1 - \tau_2 - \gamma_0 - \tau_0 - \frac{1}{2+\varepsilon_1} \tag{91}$$

Then, by Proposition 14, we have

$$\mathbb{P}_{\theta^*}\left(\frac{1}{(i+1)^{\gamma_0+\frac{1}{2+\varepsilon_1}+\delta}}\left\|J_1(\mathbf{z}(i))\right\| = 0\right) = 0 \tag{92}$$

Also, Lemma 12 implies

$$\mathbb{P}_{\theta^*}\left(\sup_{i\in\mathbb{T}_+}\left\|(I_N\otimes K)D_{\overline{H}}\mathbf{x}(i) - \mathbf{1}_N\otimes\left(\frac{1}{N}(\mathbf{1}_N\otimes I_M)(I_N\otimes K)D_{\overline{H}}\mathbf{x}(i)\right)\right\| < \infty\right) = 1 \tag{93}$$

by the boundedness of $\{\mathbf{x}(i)\}$. However, these pathwise bounds are not uniform over the sample paths and hence we use truncation arguments. For a scalar a, define its truncation $(a)^{R_0}$ at level $R_0 > 0$ by

$$(a)^{R_0} = \begin{cases} \frac{a}{|a|}\min(|a|, R_0) & \text{if } a \neq 0 \\ 0 & \text{if } a = 0 \end{cases} \tag{94}$$

For a vector, the truncation operation applies component-wise. For $R_0 > 0$, we also consider the

sequences, $\{\widehat{\mathbf{y}}_{R_0}(i)\}_{i\geq 0}$, given by

$$\begin{aligned}\widehat{\mathbf{y}}_{R_0}(i+1) &= \left(I_{NM}-\beta(i)\overline{L}\otimes I_M-P\right)\widehat{\mathbf{y}}_{R_0}(i)-\alpha(i)\left(\left[(I_N\otimes K)D_{\overline{H}}\mathbf{x}(i)-\right.\right.\\ &\quad \left.\left.\mathbf{1}_N\otimes\left(\frac{1}{N}(\mathbf{1}_N\otimes I_M)(I_N\otimes K)D_{\overline{H}}\mathbf{x}(i)\right)\right]\right)^{R_0}\\ &\quad +\alpha(i)\left([J_1(\mathbf{z}(i))]\right)^{R_0(i+1)^{\gamma_0+\frac{1}{2+\varepsilon_1}+\delta}} \end{aligned} \tag{95}$$

We will now show that for every $R_0>0$,

$$\mathbb{P}_{\theta^*}\left(\lim_{i\to\infty}(i+1)^{\tau_0}\left(\widehat{\mathbf{y}}_{R_0}(i)\right)=0\right)=1 \tag{96}$$

for τ_0 satisfying the hypothesis 85. That, this is sufficient to conclude the assertion

$$\mathbb{P}_{\theta^*}\left(\lim_{i\to\infty}(i+1)^{\tau_0}\left(\widehat{\mathbf{y}}(i)\right)=0\right)=1 \tag{97}$$

is a consequence of the following standard argument. The pathwise boundedness of the various terms imply that for every $\varepsilon_3>0$, there exists $R_{\varepsilon_3}>0$, such that,

$$\mathbb{P}_{\theta^*}\left(\sup_{i\in\mathbb{T}_+}\left\|(I_N\otimes K)D_{\overline{H}}\mathbf{x}(i)-\mathbf{1}_N\otimes\left(\frac{1}{N}(\mathbf{1}_N\otimes I_M)(I_N\otimes K)D_{\overline{H}}\mathbf{x}(i)\right)\right\|<R_{\varepsilon_3}\right)>1-\varepsilon_3 \tag{98}$$

$$\mathbb{P}_{\theta^*}\left(\sup_{i\in\mathbb{T}_+}\|J_1(\mathbf{z}(i))\|<R_{\varepsilon_3}(i+1)^{\gamma_0+\frac{1}{2+\varepsilon_1}+\delta}\right)>1-\varepsilon_3 \tag{99}$$

For (98) we use the pathwise boundedness of $\{\mathbf{x}(i)\}$ (Lemma 12), whereas, (99) holds because the a.s. convergence in Lemma 14 implies convergence in probability. Clearly, the process $\{\widehat{\mathbf{y}}(i)\}$ agrees with the process $\{\widehat{\mathbf{y}}_{R_{\varepsilon_3}}(i)\}$ on the set where both of the above events occur. By standard manipulations, it then follows, that

$$\mathbb{P}_{\theta^*}\left(\sup_{i\in\mathbb{T}_+}\left\|\widehat{\mathbf{y}}(i)-\widehat{\mathbf{y}}_{R_{\varepsilon_3}}(i)\right\|=0\right)>1-2\varepsilon_3 \tag{100}$$

The claim in eqn. (96) would then imply

$$\mathbb{P}_{\theta^*}\left(\lim_{i\to\infty}(i+1)^{\tau_0}\left(\widehat{\mathbf{y}}(i)\right)=0\right)>1-2\varepsilon_3 \tag{101}$$

We could then establish the assertion of the lemma by taking ε_3 to zero.

Hence, in the following we establish the claim in eqn. (96) for every $R_0>0$. To this end, consider the scalar process $\{\widetilde{y}_{R_0}(i)\}_{i\in\mathbb{T}_+}$ defined recursively as

$$\widetilde{y}_{R_0}(i+1)=\left\|I_{NM}-\beta(i)L(i)-P^{NM}\right\|\widetilde{y}_{R_0}(i)+NMR_0\alpha(i)+NMR_0\alpha(i)(i+1)^{\gamma_0+\frac{1}{2+\varepsilon_1}+\delta} \tag{102}$$

with initial condition $\widetilde{y}_{R_0}(0) = \|\widehat{\mathbf{y}}_{R_0}(0)\|$. Since,

$$\begin{aligned}\|\widehat{\mathbf{y}}_{R_0}(i+1)\| &= \left\|I_{NM} - \beta(i)\overline{L}\otimes I_M - P^{NM}\right\| \|\widehat{\mathbf{y}}_{R_0}(i)\| - \alpha(i)\left\|\Big(\Big[(I_N\otimes K)D_{\overline{H}}\mathbf{x}(i)\right.\\ &\quad \left.-\mathbf{1}_N\otimes\left(\frac{1}{N}(\mathbf{1}_N\otimes I_M)(I_N\otimes K)D_{\overline{H}}\mathbf{x}(i)\right)\Big]\Big)^{R_0}\right\|\\ &\quad +\alpha(i)\left\|([J_1(\mathbf{z}(i))])^{R_0(i+1)^{\gamma_0+\frac{1}{2+\varepsilon_1}+\delta}}\right\| \qquad (103)\end{aligned}$$

it follows that,

$$\|\widehat{\mathbf{y}}_{R_0}(i)\| \leq \widetilde{y}_{R_0}(i), \quad \forall i \qquad (104)$$

By Proposition 14, for i large enough, it can be shown that

$$\left\|I_{NM} - \beta(i)\overline{L}\otimes I_M - P^{NM}\right\| = 1 - \beta(i)\lambda_2(L(i)) \qquad (105)$$

We assume w.l.o.g. that the above holds for all i. We then have

$$\begin{aligned}\widetilde{y}_{R_0}(i+1) &\leq (1-\beta(i)\lambda_2(L(i)))\,\widetilde{y}_{R_0}(i) + NMR_0\alpha(i) + NMR_0\alpha(i)(i+1)^{\gamma_0+\frac{1}{2+\varepsilon_1}+\delta}\\ &\leq (1-\beta(i)\lambda_2(L(i)))\,\widetilde{y}_{R_0}(i) + 2NMR_0\alpha(i)(i+1)^{\gamma_0+\frac{1}{2+\varepsilon_1}+\delta} \qquad (106)\end{aligned}$$

The above implies

$$\widetilde{y}_{R_0}(i+1) \leq (1-\beta(i)\lambda_2(L(i)))\,(\widetilde{y}_{R_0}(i)) + 2NMR_0\frac{1}{(i+1)^{\tau_1-\gamma_0-\frac{1}{2+\varepsilon_1}-\delta}} \qquad (107)$$

Using a result from [26], we note that $\lambda_2(\overline{L}) > 0$ implies $\mathbb{E}_{\theta^*}[\lambda_2(L(i))] > 0$ (note that this equivalence is not a consequence of Jensen's inequality, as the second eigenvalue is a concave function of the graph Laplacian.) The recursion in eqn. (108) then falls under the purview of Lemmas 4,5 (see eqns. (85,91)), and we have

$$\mathbb{P}_{\theta^*}\left(\lim_{i\to\infty}(i+1)^{\tau_0}\widetilde{y}_{R_0}(i) = 0\right) = 1 \qquad (108)$$

It then follows from eqn. (104) that

$$\mathbb{P}_{\theta^*}\left(\lim_{i\to\infty}(i+1)^{\tau_0}\widehat{\mathbf{y}}_{R_0}(i) = 0\right) = 1 \qquad (109)$$

The assertion is then immediate. ■

Lemma 15 characterizes the proximity of the sensor estimates $\{\mathbf{x}_n(i)\}$ to the network averaged estimate $\{\mathbf{x}_{\text{avg}}(i)\}$. To infer the convergence of the sensor estimates to θ^*, it then suffices to study the limiting properties of $\{\mathbf{x}_{\text{avg}}(i)\}$. This is achieved in two steps. In the following, we consider the class of linear centralized estimators of the parameter θ, and establish its relation to the network averaged estimator $\{\mathbf{x}_{\text{avg}}(i)\}$. In particular, we investigate the rate at which $\{\mathbf{x}_{\text{avg}}(i)\}$ converges to the class of centralized estimators. Properties of the centralized estimators are then used to infer the convergence of $\{\mathbf{x}_{\text{avg}}(i)\}$ (and hence, that of $\{\mathbf{x}_n(i)\}$) to θ^*.

The following result is the first step towards characterizing the convergence rate of the network averaged estimator $\{\mathbf{x}_{\text{avg}}(i)\}$ to θ^*. It establishes the relation between $\{\mathbf{x}_{\text{avg}}(i)\}$ and the class of centralized estimators $\{\mathbf{u}(i)\}$ introduced in Definition 2.

Lemma 16 Let $\{\mathbf{u}(i)\}$ be the centralized estimate sequence defined in 2 with $\tau_c = \tau_1$, $a_c = a$ and $K_c = K$. Then,

(1)

$$\mathbb{P}_{\theta^*}\left(\lim_{i\to\infty}\left\|\mathbf{x}_{\text{avg}}(i) - \mathbf{u}(i)\right\| = 0\right) = 1 \tag{110}$$

(2) Let τ_0 satisfy the assumption

$$0 < \tau_0 < \tau_1 - \tau_2 - \gamma_0 - \frac{1}{2+\varepsilon_1} \tag{111}$$

Also, if $\tau_1 = 1$, assume that the constant a in assumption **(A.6)** satisfies

$$a > \frac{N\tau_0}{\lambda_{\min}(KG)} \tag{112}$$

Then,

$$\lim_{i\to\infty}(i+1)^{\tau_0}\left(\mathbf{x}_{\text{avg}}(i) - \mathbf{u}(i)\right) = 0 \tag{113}$$

Proof: We note that the averaged update may be written as

$$\begin{aligned}
\mathbf{x}_{\text{avg}}(i+1) &= \mathbf{x}_{\text{avg}}(i) + \frac{\alpha(i)}{N}K\sum_{n=1}^{N}\overline{H}_n^T\mathbf{z}_n(i) - \frac{\alpha(i)}{N}K\sum_{n=1}^{N}\overline{H}_n^T\overline{H}_n\mathbf{x}_n(i) \\
&= \mathbf{x}_{\text{avg}}(i) + \frac{\alpha(i)}{N}K\sum_{n=1}^{N}\left(\overline{H}_n^T\mathbf{z}_n(i) - \overline{H}_n^T\overline{H}_n\mathbf{x}_{\text{avg}}(i)\right) \\
&\quad -\frac{\alpha(i)}{N}K\sum_{n=1}^{N}\overline{H}_n^T\overline{H}_n\left(\mathbf{x}_n(i) - \mathbf{x}_{\text{avg}}(i)\right)
\end{aligned} \tag{114}$$

Define the process $\{\widetilde{\mathbf{u}}(i)\}$ by

$$\widetilde{\mathbf{u}}(i) = \mathbf{x}_{\text{avg}}(i) - \mathbf{u}(i) \tag{115}$$

We then have

$$\widetilde{\mathbf{u}}(i+1) = (I_M - \frac{\alpha(i)}{N}KG)\widetilde{\mathbf{u}}(i) - \frac{\alpha(i)}{N}K\sum_{n=1}^{N}\overline{H}_n^T\overline{H}_n\left(\mathbf{x}_n(i) - \mathbf{x}_{\text{avg}}(i)\right) \tag{116}$$

Now choose δ, such that,

$$0 < \delta < \tau_1 - \tau_2 - \gamma_0 - \tau_0 - \frac{1}{2+\varepsilon_1} \tag{117}$$

Since $\tau_0 + \delta < \tau_1 - \tau_2 - \gamma_0 - \frac{1}{2+\varepsilon_1}$, by Lemma 15, it follows that,

$$\mathbb{P}_{\theta^*}\left(\lim_{i\to\infty}(i+1)^{\tau_0+\delta}\left\|\sum_{n=1}^{N}\overline{H}_n^T\overline{H}_n\left(\mathbf{x}_n(i) - \mathbf{x}_{\text{avg}}(i)\right)\right\| = 0\right) = 1 \tag{118}$$

Then, there exists a finite random variable R_3, such that,

$$\mathbb{P}_{\theta^*}\left(\left\|\sum_{n=1}^{N}\overline{H}_n^T\overline{H}_n\left(\mathbf{x}_n(i)-\mathbf{x}_{\text{avg}}(i)\right)\right\|\leq R_3(i+1)^{-\tau_0-\delta}\quad\forall i\in\mathbb{T}_+\right)=1 \tag{119}$$

Note, by hypothesis, the matrix KG is symmetric and $\alpha(i)\to 0$. Hence, there exists a constant $c_{10}>0$, such that, for sufficiently large i,

$$\left\|I_M-\frac{\alpha(i)}{N}KG\right\|\leq 1-c_{10}\alpha(i)$$

Writing ω-wise and introducing another constant $c_{11}>0$, we have

$$\|\widetilde{\mathbf{u}}(i+1,\omega)\|\leq(1-c_{10}\alpha(i))\,\|\widetilde{\mathbf{u}}(i,\omega)\|+c_{11}\alpha(i)R_3(\omega)(i+1)^{-\tau_0-\delta} \tag{120}$$

for i greater than some sufficiently large $i_4(\omega)$. We then have

$$\|\widetilde{\mathbf{u}}(i+1,\omega)\|\leq(1-c_{KG}\alpha(i))\,\|\widetilde{\mathbf{u}}(i,\omega)\|+c_{11}R_3(\omega)(i+1)^{-\tau_1-\tau_0-\delta} \tag{121}$$

A pathwise (fixed ω) application of Lemma 4 and Lemma 5 and noting that the above holds for ω in a set of full measure yield the assertions. ■

VI. Proofs of main results

A. Proof of Theorem 10

Consider the first assertion. Since the $\mathcal{GLU}$ parameters (τ_1,a,τ_2,b,K) satisfy assumption (A.6), we note

$$.5+\gamma_0<\tau_1\leq 1 \tag{122}$$

Choose $\tau_c=\tau_1$ and $K_c=K$. It then follows that the centralized estimator $\{\mathbf{u}(i)\}$ (Definition 2) with design parameters is good. Hence, by Proposition 7 it is consistent, i.e.,

$$\mathbb{P}_{\theta^*}\left(\lim_{i\to\infty}\mathbf{u}(i)=\theta^*\right)=1 \tag{123}$$

Taking $\tau_0=0$ in Lemma 15, we have

$$\mathbb{P}_{\theta^*}\left(\lim_{i\to\infty}\left(\mathbf{x}(i)-\mathbf{1}_N\otimes\mathbf{x}_{\text{avg}}(i)\right)=0\right)=1 \tag{124}$$

The first assertion of Theorem 10 is then an immediate consequence of (123)-(124) and Lemma 16 (first assertion.)

The second assertion of Theorem 10 is a direct consequence of Lemma 15 and Lemma 16 (first assertion.)

B. Proof of Theorem 11

By hypothesis of Theorem 11, we have

$$\tau_1 = 1, \quad \frac{1}{2+\varepsilon_1} + \tau_2 < .5 \tag{125}$$

Hence, $\tau_1 - \tau_2 - \frac{1}{2+\varepsilon_1} > .5$. Since, $a > \frac{N}{2\lambda_{\min}(KG)}$, there exists $\varepsilon_5 > 0$, small enough, such that,

$$a > \frac{N(.5+\varepsilon_5)}{\lambda_{\min}(KG)} \tag{126}$$

By the above, we can always choose τ_0 satisfying the condition:

$$.5 < \tau_0 < \max\left(.5+\varepsilon_5, \tau_1 - \tau_2 - \frac{1}{2+\varepsilon_1}\right) \tag{127}$$

For such τ_0, we clearly have $a > \frac{N\tau_0}{\lambda_{\min}(KG)}$, and hence by Theorem 10 (second assertion), we conclude

$$\mathbb{P}_{\theta^*}\left(\lim_{i\to\infty}(i+1)^{\tau_0}\left(\mathbf{x}_n(i) - \mathbf{u}(i)\right) = 0\right) = 1 \tag{128}$$

where $\{\mathbf{u}(i)\}$ is the centralized estimator with design parameters (τ_c, a_c, K_c), such that, $a_c = a$, $\tau_c = \tau_1$, $K_c = K$. It then follows by Proposition 9, that,

$$\sqrt{(i+1)}\left(\mathbf{u}(i) - \theta^*\right) \Longrightarrow \mathcal{N}(\mathbf{0}, S_c(K)) \tag{129}$$

Since, τ_0 in (128) is strictly greater than .5, the sequences $\{\mathbf{x}_n(i)\}$ and $\{\mathbf{u}(i)\}$ are indistinguishable in $\sqrt{(i+1)}$ scale, and it can be shown using standard properties of stochastic convergence, that,

$$\sqrt{(i+1)}\left(\mathbf{x}_n(i) - \theta^*\right) \Longrightarrow \mathcal{N}(\mathbf{0}, S_c(K)) \tag{130}$$

The second assertion follows by choosing $K = K^*$ in the first.

VII. Conclusions

The paper considers gossip linear estimation of an unknown large dimensional parameter (or large scale static random field) observed by a sparsely interconnected network of sensors operating under the gossip communication protocol. We consider this problem under very general conditions on the noise assumptions and communication failures (including, link or channel failures, besides the usual measurement noise assumptions.) Due to the large scale of the field, the sensors are local, i.e., they observe only a small fraction of the field. To obtain a global estimate, the sensors need to cooperate. The class of gossip distributed linear estimators we study combines two terms: a *consensus* term that updates at each sensor its current estimate with the state estimates provided by the neighbor(s) when they gossip; and an *innovations* or *sensing* term that updates the current sensor estimate with the new observation. The linear gossip distributed estimators that we analyze exhibit a mixed time scale–one that is associated with the consensus and the other with the innovations. This forces us to develop new analytical tools to establish their asymptotic properties. This is because in gossip distributed estimation,

the innovation term is not a martingale difference process, as in previous work on mixed time scale stochastic approximation algorithms, e.g., [4]; so, a key step in our analysis is to derive pathwise strong approximation results to characterize the rate at which the innovation process converges to a martingale difference process. The paper establishes a distributed observability condition–global observability, a condition on the sensing devices, i.e., the local measurements, plus mean connectedness, a structural condition on the communication network as provided by gossip. We show that under this condition the distributed estimators performance approaches the asymptotic performance of the optimal centralized estimators, namely, the distributed estimators are consistent and asymptotically normal. This is significant, as it shows that, under reasonable assumptions, a distributed gossip estimator is as good as a centralized one, the latter having access to all sensor observations at all times. As mentioned, the distributed gossip estimator has two time scales, which involves setting two gain sequences, one for the local innovations at each sensor and the other for estimate fusion (consensus) across sensors. To design good distributed gossip estimators, these gains should be chosen properly, namely, the consensus gain should decay at a slower rate than the innovation gain. In the absence of quantization or channel noise, the paper shows that it is possible to choose the consensus weight sequence such that its squared sum goes to ∞, in contrast to the innovation weight sequence whose squared sum needs to be finite. This tuning of the different gain sequences leads to an asymptotic time scale separation, the rate of information dissemination dominating the rate of reduction of uncertainty by observation acquisition. This is not possible with quantized or noisy transmissions, as each consensus step introduces noise, preventing proper adjustment of the gain sequences. The paper interprets the fundamental convergence results on distributed gossip estimation in two interesting contexts: 1) when the observations are (conditionally) independent, the distributed estimator achieves the same performance (in terms of asymptotic variance) as the best centralized linear estimator; and 2) the maximum rate at which the observation noise power (variance) can increase with time and still the estimators to remain consistent is the same for the centralized and the gossip linear distributed estimators.

REFERENCES

[1] S. Kar, J. M. F. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: nonlinear observation models and imperfect communication," August 2008, submitted to the IEEE Transactions on Information Theory, 51 pages. [Online]. Available: http://arxiv.org/abs/0809.0009

[2] S. Kar and J. M. F. Moura, "Gossip and distributed Kalman filtering: Weak consensus under weak detectability," 2010, submitted for publication.

[3] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge, UK: Cambridge University Press, 2008.

[4] S. B. Gelfand and S. K. Mitter, "Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$," *SIAM J. Control Optim.*, vol. 29, no. 5, pp. 999–1018, September 1991.

[5] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. AC-31, no. 9, pp. 803–812, 1986.

[6] J. N. Tsitsiklis, "Problems in decentralized decision making and computation," Ph.D., Massachusetts Institute of Technology, Cambridge, MA, 1984.

[7] D. Bertsekas, J. Tsitsiklis, and M. Athans, "Convergence theories of distributed iterative processes: A survey," *Technical Report for Information and Decision Systems, Massachusetts Inst. of Technology, Cambridge, MA*, 1984.

[8] H. Kushner and G. Yin, "Asymptotic properties of distributed and communicating stochastic approximation algorithms," *Siam J. Control and Optimization*, vol. 25, no. 5, pp. 1266–1290, Sept. 1987.

[9] A. Das and M. Mesbahi, "Distributed linear parameter estimation in sensor networks based on laplacian dynamics consensus algorithm," in *3rd Annual IEEE Communications Society on Sensor and Ad Hoc Communications and Networks*, vol. 2, Reston, VA, USA, 28-28 Sept. 2006, pp. 440–449.

[10] I. D. Schizas, A. Ribeiro, and G. B. Giannakis, "Consensus in ad hoc wsns with noisy links - part i: Distributed estimation of deterministic signals," *IEEE Transactions on Signal Processing*, vol. 56, no. 1, pp. 350–364, January 2008.

[11] S. Kar, S. A. Aldosari, and J. M. F. Moura, "Topology for distributed inference on graphs," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2609–2613, June 2008.

[12] U. A. Khan and J. M. F. Moura, "Distributing the kalman filter for large-scale systems," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, p. 49194935, October 2008.

[13] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 3122–3136, July 2008.

[14] S. Stankovic, M. Stankovic, and D. Stipanovic, "Decentralized parameter estimation by consensus based stochastic approximation," in *46th IEEE Conference on Decision and Control*, New Orleans, LA, USA, 12-14 Dec. 2007, pp. 1535–1540.

[15] I. Schizas, G. Mateos, and G. Giannakis, "Stability analysis of the consensus-based distributed lms algorithm," in *Proceedings of the 33rd International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, Nevada, USA, April 1-4 2008, pp. 3289–3292.

[16] S. Ram, V. Veeravalli, and A. Nedic, "Distributed and recursive parameter estimation in parametrized linear state-space models," *to appear in IEEE Transactions on Automatic Control*.

[17] F. R. K. Chung, *Spectral Graph Theory*. Providence, RI : American Mathematical Society, 1997.

[18] B. Mohar, "The Laplacian spectrum of graphs," in *Graph Theory, Combinatorics, and Applications*, Y. Alavi, G. Chartrand, O. R. Oellermann, and A. J. Schwenk, Eds. New York: J. Wiley & Sons, 1991, vol. 2, pp. 871–898.

[19] B. Bollobas, *Modern Graph Theory*. New York, NY: Springer Verlag, 1998.

[20] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE/ACM Trans. Netw.*, vol. 14, no. SI, pp. 2508–2530, 2006.

[21] R. Subramanian and K. V. Bhagwat, "On a theorem of wigner on products of positive matrices," *Proceedings Mathematical Sciences*, vol. 88, no. 1, pp. 31–34, January 1979.

[22] M. Nevel'son and R. Has'minskii, *Stochastic Approximation and Recursive Estimation*. Providence, Rhode Island: American Mathematical Society, 1973.

[23] S. Kar and J. M. F. Moura, "Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 355–369, January 2009.

[24] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation and Time Series Analysis*. Addison Wesley, 1990.

[25] O. Kallenberg, *Foundations of Modern Probability*, 2nd ed. Springer Series in Statistics., 2002.

[26] S. Kar and J. M. F. Moura, "Sensor networks with random links: Topology design for distributed consensus," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 3315–3326, July 2008.